ASSESSING EVIDENCE IN MULTIPLE HYPOTHESES

BU-1184-M January 1993

Constantinos Goutis
University College London

George Casella[1]
Martin T. Wells
Cornell University

Key Words and Phrases: Hypothesis Testing, Decision Theory, p-values, meta-analysis.

[1] Research supported by National Science Foundation Grant No. DMS9100839 and National Securrity Agency Grant No. 90F-073

Summary

We formulate the problem of choosing between two hypotheses as a problem of constructing a data-dependent evidential measure for or against the null hypothesis. Such a measure should obey some intuitive desiderata which we state as axioms. We examine critically the behaviour of common rules of assessing evidence in higher dimensions, such as combinations of p-values and Bayes posterior probabilities, and clarify the connection between hypothesis testing and our approach. Decision theoretical tools are also used to evaluate various rules and a general admissibility result is provided. A discussion comparing different methods of assessing evidence is included.

1. Introduction

Suppose it is of interest to test the hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 \tag{1.1}$$

using observed values x of the random variable X having a distribution depending on θ. The data and the parameter may be vector valued, and we can take $\Theta_1 = \Theta_0^c$ without loss of generality. The classical decision theoretical approach to the problem of choosing between H_0 and H_1 is built around the Neyman-Pearson Lemma and the classical hypothesis tests theory, as in Lehmann (1985). Such an approach yields a $0-1$ answer, rules that are often criticised as being data insensitive. Once a partition of the sample space into two regions has been decided, all values of data in the acceptance region are considered to cast the same evidence for H_0 whereas all data in the rejection region bring the same evidence against H_0. This binary behaviour has been considered as a serious drawback from theoretical point of view (e. g. DeGroot 1973, Berger 1985, Kiefer 1977, Robinson 1979). Several authors have proposed alternative approaches to the problem including the use of Bayesian posterior probabilities (DeGroot 1973), the report of conditional confidence (Kiefer 1977) and a decision theoretical approach to the estimation of accuracy of testing (Hwang *et al.* 1992). A basic consideration in all these approaches is to provide a data dependent measure of evidence for or against the null hypothesis.

Practitioners seem to act under the same paradigm and not under the Neyman-Pearson theory. This might explain the wide-spread use of p-values, despite the non-rigorous derivation and interpretation which leads to a wide-spread misuse. Subject matter journals are flooded with p-values and their interpretation is, loosely speaking, "The smaller the p-value, the more evidence against H_0". This interpretation borrows from the Bayesian approach, which offers valid measures of plausibility of the null hypothesis such as the posterior probability or the Bayes factor. The latter quantities are based on the posterior distribution of the parameter given the data, so they are inherently data sensitive.

One might ask if decision theoretic tools, other than the traditional Neyman-Pearson-Wald approach, can help in measuring evidence in hypothesis testing. Since we are interested in estimating the viability of the set specified by H_0 we can transform the testing problem into an estimation problem. Instead of deciding whether H_0 or H_1 is true we can consider the problem of estimating the function $I(\theta \in \Theta_0)$ (where $I(\cdot)$ denotes the indicator of an event). The performance of a decision rule, $\phi(x)$, is evaluated with respect to a loss function

$$L(\theta, \phi) = d\Big(I(\theta \in \Theta_0) - \phi(x)\Big), \tag{1.2}$$

where the function d(t) is minimised at $t = 0$, nondecreasing for $t > 0$, and nonincreasing for $t < 0$.

An important point to note is that we are considering this problem as one of estimation, not of deciding between H_0 and H_1. Thus, we are making an assessment of H_0, rather than drawing a

conclusion about H_0. To assess H_0, we try to estimate $I(\theta \in \Theta_0)$ with $\phi(x)$, where we consider $I(\theta \in \Theta_0)$ the quantity of interest implicit in the setup of (1.1). The rule $\phi(x)$ has the interpretation that large values support H_0 while small values support H_1, much like a p-value or a posterior probability of H_0, and thus $\phi(x)$ can be used by an experimenter in a similar way. An immediate consequence of this formulation, including the shape of the function $L(\theta, \phi)$, is that any reasonable $\phi(x)$ must be in the interval [0, 1] for all values of the data. We will use the term *evidential statistic* or *measure of evidence* for $\phi(x)$. Note, however, that $\phi(x)$ does not measure "evidence" in the formal traditional sense, as done through the likelihood ratio (Birnbaum 1962, Royall 1986).

Although (1.1) and (1.2) define the general problem of estimation in testing hypotheses, we will only consider some special cases in what follows, using losses of the form

$$L_m(\theta, \phi) = \left| I(\theta \in \Theta_0) - \phi(X) \right|^m, \quad m = 1, 2, \tag{1.3}$$

with associated risk functions

$$R_m(\theta, \phi) = E_\theta \left| I(\theta \in \Theta_0) - \phi(X) \right|^m, \quad m = 1, 2. \tag{1.4}$$

Note that standard Neyman-Pearson-type results may be viewed as decision-theoretic results using a loss of the form (1.3) with $m = 1$. In particular, the Bayes rules with respect to $L_1(\theta, \phi)$, are Neyman-Pearson-type solutions. Note that $\phi(x)$ estimates $I(\theta \in \Theta_0)$, and is not a rejection probability, so a Neyman-Pearson *critical function* would be equivalent to $1 - \phi(x)$.

Alternatively, the Bayes rules with respect $L_2(\theta, \phi)$ are the posterior probabilities of H_0 given the data. Hence the loss $L_2(\theta, \phi)$ yields optimal solutions belonging to the interval [0, 1], conforming to our desire to give a measure of evidence of H_0 and not a $0 - 1$ answer. However, evidential rules need not be restricted to Bayes rules. One can use heuristic methods to derive a $\phi(x)$ (for example $\phi(x)$ may be a p-value). By transforming the testing problem into an estimation one, we can evaluate rules using their risk $R_2(\theta, \phi)$, although risk considerations are not the only ones for a rule to be optimal. This approach was taken in Hwang *et al.* (1992) with fruitful results for the univariate case.

The multivariate case can be treated in a similar spirit. However, for the testing setup there is a critical difference. While in the univariate case one can satisfactorily reduce the plausible procedures using optimality criteria (admissibility, consistency, invariance etc), if x and θ are vector valued the remaining procedures that are optimal (in some sense) are too numerous to choose from. For example, suppose that a multivariate test is considered as a combination of univariate tests, where the components of the vector valued data are independent of each other and the set Θ_0 is a Cartesian product of subsets of $\Re$. Then even if the choice of univariate tests is clear-cut, there are many ways to combine the results in a single statement about the truth of the joint null hypothesis. Choosing among them is usually termed as the problem of meta-analysis (Hedges and Olkin 1985).

Our goal in this paper is to formalise evidence in the multivariate case, state the appropriate optimality criteria and examine if the evidential statistics commonly used are satisfactory. In Section 2

we state the assumptions and give the appropriate definitions. In Section 3 we review the methods of constructing p-values for a multivariate hypothesis by combining individual p-values, and we derive an impartial Bayes rule. We also discuss how we would like evidential statistics to behave and give a theorem relating posterior probabilities with p-values by examining the likelihood at various points of the parameter space. In Section 4 we state formally the intuitive requirements as axioms and examine in detail the behaviour of some common evidential measures. Section 5 relates hypothesis testing with evidence in higher dimensions and examines evidence from a decision theoretical point of view whereas Section 6 contains a complete class theorem for estimators of $I(\theta \in \Theta_0)$ under the loss $L_2(\theta, \phi)$. A discussion is included in Section 7.

2. Definitions and assumptions

A common evidential measure is the p-value. However, there does not seem to exist a universally accepted definition of a p-value. In many cases it has the form $P(S(X) > s(x))$ where x is the observed value of the random variable X and S is some statistic for which large values are considered evidence against H_0. However, this is both rather vague since evidence is not a well defined term, and also ambiguous since such a statistic might not be obvious and or there might be more than one obvious statistics (see the binomial example of Berger and Delampady 1987). Often a p-value is defined via a test, but the requirement of a test seems extraneous. We will use a more general definition which accommodates all cases.

Definition 1. A statistic p(x) is a p-value if, for $\theta \in \Theta_0$, the supremal distribution of p(X) is uniform (0, 1), that is, if

$$\sup_{\theta \in \Theta_0} P(\, p(X) \leq u) = u, \qquad 0 \leq u \leq 1. \tag{2.1}$$

If the set Θ_0 has more than one element the distribution of p(X) could be different for various elements of Θ_0. However in most cases of interest there is a rather unambiguous "extreme boundary" point of Θ_0 such that p(X) is uniform if the parameter is this point and stochastically larger than a uniform for all other points of Θ_0. Furthermore p(X) is often stochastically smaller than a uniform for $\theta \in \Theta_1$, but the definition does not require the specification of an alternative hypothesis. The p-value often has the form $1 - F_{\theta_0}(S(x))$ where θ_0 is on the boundary of the null hypothesis and F_θ is the cumulative distribution function of a statistic S when the true parameter is θ.

Note that Definition 1 neither implies that p(x) is optimal or sensible in any sense, nor gives any recipe as to how to construct a p-value. Similar to the definitions of other statistical procedures (e. g. estimators or tests), it simply gives a criterion to judge if a function of the data is a p-value for a particular null hypothesis.

Example 1. Consider that $X_i \sim N(\theta, 1)$, $i = 1, 2, \ldots, n$, independently and we wish to test

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0 . \tag{2.2}$$

From sufficiency considerations, it turns out that a reasonable statistics to use is $\overline{X} = \sum_{i=1}^{n} X_i$ and large values of $\overline{X}$ indicate large values of θ. Since the set $\Theta_0 = (-\infty, \theta_0\ ]$ contains small values of θ, quantities of the form

$$P_\theta(\overline{X} > \overline{x}) = 1 - \Phi((\overline{x} - \theta)/\sqrt{n}), \qquad \theta \leq \theta_0, \tag{2.3}$$

where Φ is the standard normal cumulative distribution function, might be useful in defining a p-value. It is straightforward to see that $P_\theta(\overline{X} > \overline{x})$ is increasing in θ so the most conservative (largest) value of (2.3) is attained at θ_0. Hence one can define as p-value the statistic

$$p(\overline{x}) = 1 - \Phi((\overline{x} - \theta_0)/\sqrt{n}). \tag{2.4}$$

For $\theta = \theta_0$, $p(\overline{X})$ is exactly the one minus the probability integral transform, so if the true parameter equals θ_0 then it has a uniform distribution. Since (2.4) is decreasing in $\overline{x}$ and the normal distribution has the monotone likelihood ratio property, smaller values of θ generate larger values of $p(\overline{X})$. Hence for $\theta < \theta_0$, $p(\overline{X})$ is stochastically larger than a uniform random variable, whereas for $\theta > \theta_0$, it is stochastically smaller.

It turns out that a p-value is closely related to a test, or to be more accurate, to a family of testing procedures, one for each α level. The following theorem is a generalisation of Lehmann (1985) p. 170 and shows the relation.

Theorem 1. For the hypothesis in (1.1), suppose that R_α is the rejection region of a test of size α. Furthermore, suppose that R_α is defined for every α and

$$\alpha_1 \geq \alpha_2 \Rightarrow R_{\alpha_2} \subseteq R_{\alpha_1} . \tag{2.5}$$

For every x, define

$$p(x) = \inf \{\alpha : x \in R_\alpha\} \tag{2.6}$$

then p(x) is a p-value according to the Definition 1.

Proof. It suffices to show that for every $\alpha_0 \in [0, 1]$

$$\sup_{\theta \in \Theta_0} P_\theta(p(X) \leq \alpha_0) = \alpha_0 . \tag{2.7}$$

Note that $p(X) \leq \alpha_0$ is equivalent to $X \in R_\alpha$ for every $\alpha \geq \alpha_0$ by (2.5) and (2.6). Hence,

$$\begin{aligned} \sup_{\theta \in \Theta_0} P_\theta(p(X) \leq \alpha_0) &= \sup_{\theta \in \Theta_0} P_\theta(\, X \in R_\alpha \quad \forall \quad \alpha \geq \alpha_0) \\ &= \sup_{\theta \in \Theta_0} P_\theta(\, X \in \bigcap_{\alpha \geq \alpha_0} R_\alpha) \end{aligned}$$

$$= \sup_{\theta \in \Theta_0} P_\theta(X \in R_{\alpha_0})$$
$$= \alpha_0$$

establishing (2.7) . □

Conversely, from a p-value p(x) one can almost trivially define a test of size α by the rejection region

$$R_\alpha = \{ x : p(x) \leq \alpha \} . \tag{2.8}$$

Indeed p-values are often used to derive a test by rejecting the null hypothesis when the p-value is smaller than a particular level α. Obtaining the correct rejection probability for the test is major consideration for requiring the p-values to be uniform in H_0. Note that the only restriction is that the rejection regions are nested in the sense of (2.5), which seems a minimal requirement for a sensible test. Apart from that, a test corresponding to a p-value need not be an optimal or sensible procedure, reflecting the fact that the definition of p-value is quite general. It is also wrong to assume that any optimality of the test is automatically transferred to the p-value or vice versa.

We now focus our attention on the multivariate problem, where things are a bit more complicated. Suppose that $X_i \sim f(x_i \mid \theta_i)$ independently, for i = 1, 2, ..., k. To avoid trivialities we assume that the support of $f(x \mid \theta_i)$ does not depend on θ_i. The observations X_i are assumed one dimensional, perhaps after reduction by sufficiency or other considerations. We will denote the data $(x_1, \ldots, x_k)$ by $\mathbf{x}$ whereas θ will be the vector $(\theta_1, \ldots, \theta_k)$. The data x_i are obtained from separate experiments providing separate pieces of evidence about the hypotheses

$$H_{0i} : \theta_i \in \Theta_{0i} \quad \text{vs.} \quad H_{1i} : \theta_i \notin \Theta_{0i}. \tag{2.9}$$

The most important cases are $\Theta_{0i} = (-\infty, \theta_{0i}]$ or $\Theta_{0i} = [\, \underline{\theta}_{0i}, \overline{\theta}_{0i} \,]$, where $\underline{\theta}_{0i} \leq \overline{\theta}_{0i}$, that is, the one-sided and two-sided hypotheses. Without loss of generality we can take $\theta_{0i} = 0$, $\underline{\theta}_{0i} = -\epsilon$, $\overline{\theta}_{0i} = \epsilon$, for $\epsilon \geq 0$. It might be of interest to combine the evidence from all x_i into a single statement about the truth of the combined null hypothesis $H_0 = \bigcap_i H_{0i}$, in which case the test of interest will have the form (1.1) with $\Theta_0 = (-\infty, 0]^k$ or $\Theta_0 = [\, -\epsilon, \ \epsilon \,]^k$, i. e.

$$H_0 : \ \theta_i \leq 0, \ i = 1, \ldots, k \quad \text{vs.} \quad H_1 : \ \theta_i > 0, \ \text{for some i} \tag{2.10}$$

and

$$H_0 : \ |\theta_i| \leq \epsilon, \ i = 1, \ldots, k \quad \text{vs.} \quad H_1 : \ |\theta_i| > \epsilon, \ \text{for some i.} \tag{2.11}$$

For specific distributions one can view the problem of testing (2.10) or (2.11) as a special case of (1.1) and derive a single appropriate test. However we view the problem of combining evidence in a general setup where the form of the distribution of X_i is not used, either because it is unknown or because the values of x_i are unknown. Hence the procedures that we will examine are "omnibus" in that they are not distribution specific. We will assume that for each individual hypothesis of (2.9) we

have a p-value available, which we will denote by $p(x_i)$ or p_i.

The problem of combining evidence using a combination of p-values goes back at least to Tippett (1931) and Fisher (1932). Other important references include Birnbaum (1954) and Hedges and Olkin (1985). It is worth noting that, given a testing procedure and the density $f(\cdot \mid \theta)$, we can sometimes reconstruct the values of x_i (or at least the information from x_i that is used in the test) from the individual p-values $p(x_i)$ and vice versa.

To avoid detailing the technical assumptions in each statement in the paper we will assume that the densities $f(x_i \mid \theta_i)$ have the monotone likelihood ratio property and the parameterisation is such that large values of x_i indicate large values of θ_i. Hence if $\theta_i \notin \Theta_{0i}$, the $p(X_i)$ will be stochastically smaller than a uniform random variable (Lehmann 1985 p. 170). Note that though it is not a requirement of Definition 1, in practice all p-values that are used satisfy it. It is also worth noting that for (2.10) and (2.11) there are points in the parameter space that can be considered "extreme boundary" points between the null and the alternatives. When we talk about *extreme boundary points* or *extreme points of the null* we will refer to $\theta = (0, 0, \ldots, 0)$ for (2.10) and $\theta = (\pm\epsilon, \pm\epsilon, \ldots, \pm\epsilon)$ for (2.11).

3. Measuring Evidence

In this section we review some of the more common rules used to measure evidence, in particular the p-values discussed by Marden (1991). Also, we look at a (generalized) Bayes rule using a sequence of priors that concentrate mass 1/2 on H_0 and 1/2 on H_1. We examine the behavior of these rules as measures of evidence, and formulate a set of axioms that evidential measures should satisfy. We will state everything in term of H_0 and H_1 given by (2.10) but the translation to (2.11) is straightforward. If there is a substantial difference it will be explicitly stated.

3.1 Combining p-values

As seen in Section 2, the defining characteristic of a p-value is its uniformly distribution under H_0. Given any statistic $S = S(X_1, \ldots, X_k)$, we can construct a p-value based on observing $S = s$ as

$$p_S(s) = \sup_{\theta \in \Theta_0} P_\theta(S \geq s). \tag{3.1}$$

If S has a continuous distribution, then under H_0 the random variable $p_S(S)$ will be stochastically greater than a uniform (0, 1) random variable and typically for some value $\theta \in \Theta_0$ it will be uniformly distributed. If S is discrete $p_S(S)$ will be stochastically greater than a uniform random variable. For $p_S(s)$ to measure evidence against (or for) H_0, its behavior (or equivalently the behavior of S) on H_1 must also be examined. This is where our rules for measuring evidence become important.

Two popular methods for creating a p-value for the hypothesis (1.1) using individual p-values of

(2.9) are based on ideas of Fisher (1932) and Tippett (1931). (Tippett's rule is actually a special case of that of Wilkinson (1951)). Other rules are the normal, the sum, Pearson's and maximum. Table 1 shows the statistic S used in each case as function of $p(x_i)$. Note that for each rule, S is unique up to one-to-one transformations. For more details about combining p-values to construct an overall p_S see Birnbaum (1954) and Hedges and Olkin (1985).

Table 1: Methods for combining p-values

Method	Statistic
Fisher	$\prod_{i=1}^{k} p_i$
Tippett	$\min_{1 \le i \le k} p_i$
Normal	$\sum_{i=1}^{k} \Phi^{-1}(p_i)$
Sum	$\sum_{i=1}^{k} p_i$
Pearson	$1 - \prod_{i=1}^{k} (1 - p_i)$
Maximum	$\max_{1 \le i \le k} p_i$

We denote the resulting p-values by p_F, p_T, p_N etc., that is

$$p_F(x) = P\left(\prod_{i=1}^{k} U_i > \prod_{i=1}^{k} p(x_i) \right) \tag{3.2}$$

$$p_T(x) = P\left(\min_{1 \le i \le k} U_i > \min_{1 \le i \le k} p(x_i) \right) \tag{3.3}$$

etc, where U_i are independent uniform (0, 1) random variables. It is easy to see all of these random variables have uniform distributions under H_0 (or, more accurately, when the individual p-values have a uniform distribution). Note that we can also use at least some of the statistics S(x) of Table 1 themselves as evidential statistics, in the sense that their value gives evidence for or against the null hypothesis. Their calibration, however, is not clear. In the future we will refer to Fisher, Tippett, Pearson and maximum statistics as well as a normalised sum statistic $\frac{1}{k}\sum p_i$. The normalisation is necessary so that it lies in the interval [0, 1]. We will not use the normal statistic since it is not clear how one could scale it to be comparable with the rest of the evidential statistics. The behaviour of these rules at different points of the null and the alternative hypothesis, however, is a criterion as to whether they are reasonable measures of evidence.

3.2 An Impartial Bayes Rule

In order to obtain some intuition about sensible evidential rules, we turn to a Bayesian derivation to see if the behavior of a Bayes rule will coincide with our thoughts on proper evidential behavior. For the hypothesis (1.1) suppose now we observe $X_i \sim f(x_i \mid \theta_i)$, and we take a prior for the θ_i' s. By considering the null hypothesis of (1.1) as an intersection of the hypotheses in (2.9) we note that the posterior probability of Θ_0, which is Bayes rule for testing (1.1) under the squared loss $L_2(\theta, \phi)$, is too small even if the prior probabilities for each H_{0i} are reasonable. The reason for that is that the prior probability of $\Theta_0 = \bigcap_i \Theta_{0i}$ is significantly smaller than the prior probability of Θ_{0i}. The phenomenon becomes stronger as the dimension k becomes larger. This may not be a problem *per se*, but we feel that the dimension induces an undesirable bias against H_0.

We can eliminate such a bias by adjusting the prior distribution so that the prior probability of H_0 and H_1 are both 1/2. We will illustrate the method for the one sided hypothesis (2.10) but the two sided case is analogous. For a given prior $\pi(\theta_1, \ldots, \theta_k)$, define

$$\gamma = \int_{\Theta_0} \pi(\theta_1, \ldots, \theta_k)\, d\theta_1, \ldots, d\theta_k \tag{3.4}$$

and take as the prior on $(\theta_1, \ldots, \theta_k)$ to be

$$\pi^*(\theta_1, \ldots, \theta_k) = \left[\frac{I(\theta \in \Theta_0)}{2\gamma} + \frac{I(\theta \notin \Theta_0)}{2(1-\gamma)}\right] \pi(\theta_1, \ldots, \theta_k). \tag{3.5}$$

Using the loss function $L_2(\theta, \phi)$ it is straightforward to calculate the Bayes rule (the posterior expectation of $I(\theta \in \Theta_0)$) as

$$\phi_\pi(x) = P_{\pi^*}(\theta \in \Theta_0 \mid x) = \frac{P_\pi(\Theta_0 \mid x)}{P_\pi(\Theta_0 \mid x) + \frac{\gamma}{1-\gamma} P_\pi(\Theta_0^c \mid x)} \tag{3.6}$$

where $P_{\pi^*}(\cdot \mid x)$ and $P_\pi(\cdot \mid x)$ are posterior probabilities under π^* and π, respectively.

If the original prior π is symmetric, then $\gamma = 2^{-k}$. Specialising to the case where $X_i \sim N(\theta_i, \sigma^2)$ and $\theta_i \sim N(0, \tau^2)$, all independent, we have

$$P_\pi(\Theta_0 \mid x) = \prod_{i=1}^{k} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi v^2}} \exp\{-\frac{1}{2v^2}(\theta_i - \delta(x_i))^2\}\, d\theta_i, \tag{3.7}$$

where $\delta(x_i) = \tau^2 x_i/(\sigma^2 + \tau^2)$ and $v^2 = \sigma^2\tau^2/(\sigma^2 + \tau^2)$. Letting $\tau^2 \to \infty$ we obtain the limiting Bayes rule

$$\phi_L(x) = \frac{\prod_{i=1}^{k} p(x_i)}{\prod_{i=1}^{k} p(x_i) + \frac{1}{2^k - 1}\left(1 - \prod_{i=1}^{k} p(x_i)\right)}, \tag{3.8}$$

where $p(x_i)$ are the individual p-values.

Examination of $\phi_L(\mathbf{x})$ yields some interesting facts. For example, $\phi_L(\mathbf{x})$ is larger than many of the evidential statistics of Table 1, such as $\prod p(x_i)$ or $\min_i p(x_i)$. However, it is smaller than the p-values based on these statistics. Thus, if one uses a statistic such as $\prod p(x_i)$ to assess evidence against H_0, this corresponds to putting prior mass less than 1/2 on H_0. (In fact, $\prod p(x_i)$ is a limiting posterior probability using a normal prior, hence weighting H_0 by 2^{-k}.) However, $\prod p(x_i)$ is smaller than most of the evidential p-values, such as p_F and p_T.

3.3 Some informal evidential desiderata

While choosing evidential rules one would want them to have some properties so that they conform to our intuition, hence we now examine and formalise such properties. In measuring evidence against H_0, we would want such evidence to increase as the parameter moves further from H_0. One way of quantifying this is to require an evidential statistic $\phi(\mathbf{x})$ to satisfy

$$\lim_{\theta_i \to \infty \; i=1,\ldots,k} E_\theta \phi(\mathbf{X}) = 0, \tag{3.9}$$

This condition merely states that if each θ_i is infinitely far from H_0, then evidence against H_0 is maximized. A stronger but not unreasonable requirement is that

$$\text{For every } i = 1,\ldots,k \qquad \lim_{\theta_i \to \infty} E_\theta \phi(\mathbf{X}) = 0. \tag{3.10}$$

Other, reasonably self-evident behavior of an evidential measure concerns monotonicity and symmetry. In the exchangeable case, it seems clear that evidence should be equal whether $\theta = (\theta_1, \theta_2, \ldots, \theta_i, \theta_j, \ldots, \theta_k)$ or $\theta = (\theta_1, \theta_2, \ldots, \theta_j, \theta_i, \ldots, \theta_k)$, that is, we require symmetry of $E_\theta \phi(\mathbf{X})$ in the arguments. Moreover, we desire our evidential statistic to be monotone. Since we want small values of $\phi(\mathbf{x})$ to signify more evidence against H_0, we require

$$E_\theta \phi(\mathbf{X}) \downarrow \theta_i, \; i = 1,\ldots,k. \tag{3.11}$$

In the two sided case, a modification of (3.11) could be

$$E_\theta \phi(\mathbf{X}) \downarrow \theta_i, \; i = 1,\ldots,k \quad \text{for } \theta_i \geq C_i \tag{3.12}$$

$$E_\theta \phi(\mathbf{X}) \uparrow \theta_i, \; i = 1,\ldots,k \quad \text{for } \theta_i \leq C_i \tag{3.13}$$

for some C_i, $-\epsilon \leq C_i \leq \epsilon$. Somewhat stricter properties are that $\phi(\mathbf{x})$ itself is symmetric in its arguments and monotone in each x_i. These imply the corresponding properties of $E_\theta \phi(\mathbf{X})$.

It is more complicated to deduce reasonable evidential behavior along an $H_0 - H_1$ boundary, for example on the set $\{\theta: \; \theta_i \leq 0 \;\; i = 1,\ldots, k-1, \; \theta_k \in \Re\}$. Also, the exact calibration of evidential statistics is unclear. More precisely, the calibration

$$\lim_{\theta_i \to -\infty \; i=1,\ldots,k} E_\theta \phi(\mathbf{X}) = 1, \tag{3.14}$$

$$E_0\, \phi(\mathbf{X}) \;=\; 1/2 \tag{3.15}$$

together with (3.11) seems reasonable for the one sided case. We might ask what class of rules satisfy (3.14) and (3.15), and see if such a requirement is reasonable and we will do so in Section 4.2. The requirement of (3.15) seems to be crucial in terms of our perception of evidence. At $\theta = (0, \ldots, 0)$, an evidential measure should be "impartial", and this concept seems to translate numerically into a statement such as (3.15). Note that for the two sided case the requirement

$$E_C\, \phi(\mathbf{X}) \;=\; 1, \tag{3.16}$$

where $C = (C_1, C_2, \ldots, C_k) \in [\, -\epsilon,\ \epsilon\,]^k$ is too strong, especially if $\epsilon = 0$. This is because, unlike the one-sided case, the set Θ_0 is compact hence, in some sense, one cannot be far from the boundary. This is specially true if $\epsilon = 0$, in which case the set Θ_0 has an empty interior. So for testing (2.11) we are content with (3.11) and

$$E_B\, \phi(\mathbf{X}) \;=\; 1/2 \tag{3.17}$$

where $B = (B_1, B_2, \ldots, B_k) \in \{\, -\epsilon,\ \epsilon\, \}^k$.

The evidential behavior that is more difficult to quantify is that which occurs along a boundary. For example, consider the configuration where $\theta_i = 0$, $i = 1, \ldots, k-1$ and $\theta_k \to -\infty$. In words, $k-1$ components are on the boundary and one component is overwhelmingly for H_0. An informal assessment of desired behavior would lead one to require the evidential statistic to "drop a dimension", that is, to assess an evidence measure of 1 for θ_k, and to behave as if we now have a $k-1$ dimensional boundary.

One reason for examining $\phi_L(\mathbf{x})$ of (3.8) was to understand its behavior on boundaries, in particular, what happens as some $\theta_i \to \infty$ while others remain fixed. To simplify calculations, we will examine the behavior of $\phi_L(\mathbf{x})$ in terms of the x_i' s which, due to the assumption of monotone likelihood ratio, is equivalent to examining behavior in terms of the θ_i' s. Suppose that, from $x_1, \ldots, x_k$, m ($\leq k$) of them are equal to 0 and the rest $k-m$ are at $-\infty$. Then, for this configuration

$$\phi_L(\mathbf{x}) \;=\; \frac{2^k - 1}{(2^k - 1) + (2^m - 1)}\,. \tag{3.18}$$

For example, if $k = 3$ and $(x_1, x_2, x_3) = (0, 0, -\infty)$, then $\phi_L = 7/10$, while if $(x_1, x_2, x_3) = (0, -\infty, -\infty)$ then $\phi_L = 7/8 > 7/10$. Thus, evidence for H_0 increases as the number of coordinates at $-\infty$ increases, quite reasonable behavior, and one that is shared by Fisher's procedure (but not Tippett's).

3.4 Extremal Behavior of Evidential Statistics

In the preceding section it was stated that an evidential statistic should approach 0 or 1 in expectation as the parameter moves deeper into H_1 or H_0, respectively. Furthermore we know that if the evidential statistic is itself a p-value then it can be arbitrarily close to 0 or 1 and it often behaves

this way if the the sample x indicates strong evidence for or against the null hypothesis. The desire to approach the limits is based mainly on common usage of p-values, and also on the interpretation of $P(\Theta_0 \mid x)$ as a measure of evidence. In this section we quantify this behavior precisely, and determine a set of conditions that guarantee the attainment of the bounds 0 and 1 by a posterior probability.

If $P(\Theta_0 \mid x)$ can range from 0 to 1, then there is a chance of Bayesian/frequentist evidence reconciliation in the sense that there may exist an (improper) prior for which $P(\Theta_0 \mid x) = p(x)$. If $P(\Theta_0 \mid x)$ is strictly between 0 and 1, then there is no chance for reconciliation. If we examine the behavior of the likelihood $f(x \mid \theta)$, we can get some idea of when reconciliation is possible. Essentially, if there exist values of θ in Θ_0 and Θ_0^c for which the likelihood can be driven to zero, then reconciliation is usually not possible. This is formalized in the following theorem, which is based on a lemma that describes a type of monotone likelihood ratio behavior.

Theorem 2. Under the conditions of Lemma A.1 (Appendix), if for every a, $b \in \overline{\mathfrak{X}}$ (the closure of the sample space) and $\theta_2 \in \Theta$ either part i) or part ii) of the Lemma fails to hold, then the p-value cannot be a posterior probability (for a prior yielding $m_\pi(x) < \infty$).

The proof follows directly from Lemma A.1. Though the conditions of the lemma seem unwieldy, they are just a detailing of obvious requirement. The following examples show this.

Examples for Theorem 2

i) Normal, one-sided. For $X \sim N(\theta, 1)$ and $\Theta_0 = (-\infty, 0]$, take $\theta_2 = 0$, $a = \infty$ and $b = -\infty$. Then Theorem 2 holds and any posterior probability ranges from 0 to 1.

ii) Normal, two-sided. For $X \sim N(\theta, 1)$ and $\Theta_0 = \{0\}$, part ii) of Lemma A.1 does not hold. Thus the posterior probability cannot reach 1.

iii) Exponential, one-sided. For $X \sim f(x \mid \theta) = \theta e^{-\theta x}$ and $\Theta_0 = (0, 1]$, part i), but not part ii), of Lemma A.1 holds (there is no b).

vi) Binomial, one-sided. For $x \sim$ Binomial(n, θ) and $\Theta_0 = [0, \theta_0]$, neither part i) nor part ii) holds. However, for the prior $\pi(\theta) = 1/\theta$ we have $\lim_{x \to 0} P(\Theta_0 \mid x) = 1$ and $P(\Theta_0 \mid x)$ is equal to the p-value. There is not contradiction as for this prior $m_\pi(0) = \infty$.

4. Formal Axioms of Evidence

We are now in the position to state formally the axioms of evidence that we would like a reasonable evidential statistic to satisfy. Then we can examine how these axioms can restrict the class of p-values (as defined in definition 1 or (3.1)) or other evidential statistic to the ones that are acceptable.

4.1 Statement of the Axioms

The axioms are stated in terms of both statistics and parameters. We first state them for the one-sided case and then we modify them for the two sided case, following the discussion in Section 3.3. For testing

$$H_0 : \theta_i \leq 0, \ i = 1, \ldots, k \quad \text{vs.} \quad H_1 : \theta_i > 0, \ \text{for some i} \tag{4.1}$$

a set of reasonable axioms for a measure of evidence $\phi(\mathbf{X})$ is:

(A1) $E_\theta \, \phi(\mathbf{X}) = 1$ if $\theta_i = -\infty$, for all $i = 1, \ldots, k$

(A2) $E_\theta \, \phi(\mathbf{X}) = 1/2$ if $\theta_i = 0$, for all $i = 1, \ldots, k$

(A3) $E_\theta \, \phi(\mathbf{X}) = 0$ if $\theta_i = +\infty$, for any $i = 1, \ldots, k$

(A4) $\phi(x_1, \ldots, x_k)$ decreasing in x_i for any $i = 1, \ldots, k$.

The values $\theta_i = -\infty$ and $\theta_i = +\infty$ are to be interpreted as the lower and upper limiting values of θ_i respectively, which may or may not be attained. Note that, since the evidential statistics belong to [0, 1], axioms (A1) and (A3) are equivalent to convergence in probability of $\phi(\mathbf{X})$ to 1 and 0 respectively.

Under the monotone likelihood ratio property an immediate consequence of (A4) is

(B1) $E_{(\theta_1, \ldots, \theta_k)} \, \phi(\mathbf{X})$ decreasing in θ_i for any $i = 1, \ldots, k$.

If the variables are exchangeable (and in the case of Bayes rules the prior is symmetric in θ_i's) it would be reasonable to expect that

(B2) $\phi(x_1, \ldots, x_i, x_j, \ldots, x_k) = \phi(x_1, \ldots, x_j, x_i, \ldots, x_k)$ for any $i, j = 1, \ldots, k$.

However, in the general case the symmetry requirement is too stringent and seems unreasonable if the random variables X_i have different distributions. Hence, in the future we will not deal with (B2).

For testing

$$H_0 : \ |\theta_i| \leq \epsilon, \ i = 1, \ldots, k \quad \text{vs.} \quad H_1 : \ |\theta_i| > \epsilon, \ \text{for some i} \tag{4.2}$$

the axioms are modified as follows:

(C1)	$E_\theta\ \phi(\mathbf{X}) \leq 1$	if $\theta_i \in [-\epsilon, \epsilon]$,	for all i = 1, . . . , k
(C2)	$E_\theta\ \phi(\mathbf{X}) = 1/2$	if $\theta_i = \pm\epsilon$,	for all i = 1, . . . , k
(C3)	$E_\theta\ \phi(\mathbf{X}) = 0$	if $\lvert\theta_i\rvert = +\infty$,	for any i = 1, . . . , k
(C4)	$\phi(x_1, \ldots, x_k)$	decreasing in x_i for $x_i \geq x_{0i}$ increasing in x_i for $x_i \leq x_{0i}$	for any i = 1, . . . , k.

Axiom (C1) is a weaker version of (A1) and is almost always satisfied. It is included for completeness. The equivalent of (B2) is now

(D1)	$E_{(\theta_1, \ldots, \theta_k)}\ \phi(\mathbf{X})$	decreasing in θ_i for $\theta_i \geq C_i$ increasing in θ_i for $\theta_i \leq C_i$, $\lvert C_i\rvert \leq \epsilon$,	for any i = 1, . . . , k

though it does not follow from (C4) immediately.

We would like a satisfactory evidential rule to satisfy axioms (A1) – (A4) or (C1) – (C4). The reasonableness of the axioms is not ad hoc, in the sense that it is a combination of intuitive thinking and examination of the behaviour of typical rules. In the following we study the p-values of Table 1, the impartial Bayes rule developed in Section 3.2 and other evidential statistics to judge if both the axioms and the rules of evidence conform to our intuition and practice. We will proceed to verify the axioms (A1) – (A4). The verification of the equivalent (C1) – (C4) for the two sided case is analogous.

4.2 Verification of the Axioms

Axiom (A1) is satisfied for all combined p-values obtained by the statistics given in Table 1, as long as the individual p-values tend to 1 in probability, i. e. if

$$\lim_{\theta_i \to -\infty} P(p(X_i) > t) = 1 \qquad \text{for every } t < 1. \tag{4.3}$$

This is also true for the statistics of Table 1 (excluding the normal and scaling them to be between 0 and 1) if they are considered as evidential measures. The Bayes rule $\phi_L(\mathbf{x})$ given by (3.8) also satisfies (A1) when (4.3) holds. However it is worth noting that (4.3) is a rather strong requirement. It is not satisfied in general in the two sided case, since, as we mentioned earlier if Θ_0 is compact, it is not possible to go deep into the null. It is satisfied however for the one sided location or scale family when the test concerns the location or scale parameter, respectively.

Axiom (A2) is automatically satisfied by all p-values, as long the individual p-values have a uniform distribution. This is true at the boundary of the null hypothesis under the monotone likelihood ratio assumption. The evidential statistics are badly calibrated with respect to (A2). The product and the minimum of the p-values have an expectation less that 1/2 whereas the maximum and Pearson statistics overestimate the evidence for the null hypothesis. This becomes more extreme as the dimension k gets larger. However, the (normalised) sum of p-values has an expectation equal to 1/2.

The Bayes rule $\phi_L(\mathbf{x})$ given by (3.8) deserves special attention, since, as we saw in Section 3.2, it was derived so that it satisfies our intuitive requirement of impartiality, hence we would like to see if it satisfies (A2). However, an application of Jensen's inequality shows that

$$E_0\, \phi_L(\mathbf{X}) \leq 1/2 \tag{4.4}$$

with equality for $k = 1$ and strict inequality for $k > 1$. This means that the evidence from $\phi_L(\mathbf{x})$ is mostly against the null hypothesis, despite the fact it was constructed so that the prior probability of the null hypothesis is 1/2. In order to obtain impartiality in the sense of (A2) we must give a significant prior weight to the null hypothesis. Table 2 shows the prior probability for selected dimensions such that if the prior π^* is constructed as (3.5) the expectation of the limiting posterior probability (3.8) equals 1/2.

Table 2: Prior probability of H_0 such that the expectation of the Bayes rule equals 0.5.

k	$P(\theta \in \Theta_0)$
2	0.612
3	0.683
4	0.741
6	0.837
10	0.943
15	0.987
20	0.997

These results are in accordance with Moreno and Cano (1989) who examined the point null hypothesis in higher dimensions and found that p-values are typically larger than posterior probabilities. The equality for $k = 1$ should be expected, in view of the results of Casella and Berger (1987) on reconciling posterior probabilities and p-values in the one-sided one-dimensional case.

Verification of axiom (A3) is more subtle. Similarly to (A1) we see that if

$$\lim_{\theta_i \to \infty} P(p(X_i) < t) = 1 \qquad \text{for every } t > 0, \tag{4.5}$$

then one can show that for all evidential statistics and the respective p-values of Table 1 as well the impartial Bayes rule $\phi_L(\mathbf{x})$, the limit of the expectation is 0 if all θ_i tend to $+\infty$. In contrast to (4.3), (4.5) is often satisfied because Θ_0^c is typically unbounded and large values of θ_i yield small $p(x_i)$ for the most commonly used p-values. However to verify axiom (A3) we must see what happens when some, but not all individual θ_i tend to $+\infty$ (hence the respective p-values tend to 0).

Note that the p-values p_F, p_N, p_S and p_P have the form

$$P\Big(\sum_{i=1}^{k} F^{-1}(U_i) \geq \sum_{i=1}^{k} F^{-1}(1 - p_i)\Big) \tag{4.6}$$

where U_i are independent uniform (0, 1) random variables and F^{-1} is the inverse of a cumulative distribution function F. To examine the limit of (4.6), we examine the behaviour of F.

For Pearson's and sum p-values we have $F(x) = 1$ for a finite x, hence (4.6) is equal to zero if and only if

$$\sum_{i=1}^{k} F^{-1}(1-p_i) = k\, F^{-1}(1) \tag{4.7}$$

that is, (4.6) is positive if at least one individual p-value is not equal to zero. A similar remark can be made about the p-value derived from $\max_i p_i$. Hence for p_P, p_S and p_M (A3) is not satisfied.

If for any $x < \infty$, we have $F(x) < 1$ (that is if the support of F is not bounded above) then one can see

$$\lim_{p_i \to 0,\ p_j \le U < 1 \text{ for } j \neq i} \quad \sum_{i=1}^{k} F^{-1}(1-p_i) = +\infty, \tag{4.8}$$

which in turn implies that (4.6) goes to zero. That means that for Fisher's and normal p-values, (A3) holds, as long as at least one individual p-value tends to zero and the other ones are bounded away from 1. Translating into parameter values, (4.6) will tend to zero, as long as at least one θ_i tends to $+\infty$ but the other ones are not too far away from the boundary of H_0 and H_1. This is somewhat unsatisfactory since it means that the individual pieces evidence can "cancel out". In other words,

$$\lim_{\theta_1 \to +\infty,\ \theta_2 \to -\infty} E_\theta\, p_N(X) \tag{4.9}$$

is indeterminate. Its value depends on the rate at which θ_1 and θ_2 go to their respective limits.

Fortunately, for Fisher's p-value there is a $x > -\infty$ such that $F(x) = 0$ (the support of F is bounded below), hence we can obtain the stronger property

$$\lim_{p_i \to 0} \quad \sum_{i=1}^{k} F^{-1}(1-p_i) = -\infty, \tag{4.10}$$

irrespective of the values of p_j, $j \neq i$, that is, if there exists strong evidence against one H_{0i}, then, irrespective of the evidence for or against other H_{0i}'s, one draws the right conclusion for $\bigcap_i H_{0i}$. Similarly, Tippett's p-value, the Fisher and Tippett statistics and the Bayes rule $\phi_L(x)$ satisfy (A3) even if some individual p_i's give strong evidence for the null hypothesis.

5. Multivariate Evidence

In the previous sections we examined the properties of evidential measures without relating them with the formal hypothesis testing setup. However, Theorem 1 guarantees an equivalence between p-values and tests so it is reasonable to expect that the implied relation will give us further insight as to how the properties of a test translate to the behaviour of the p-value as an evidential measure. In this section we relate the formal definition of a p-value in a higher dimensional setup with the respective tests and their acceptance regions. We will also try to clarify the connection between test statistics,

acceptance regions and the rules of combinations. These are straightforward in the one dimensional case, but not if the data and the parameters are vector valued. However keeping the geometric pictures in mind is always helpful.

Suppose that, as in Section 1, we test (2.10) or (2.11) and for any α we have a testing rule of the form "Reject H_0 if $\mathbf{x} \in R_\alpha$" where $R_\alpha \subseteq \Re^k$ is a rejection region such that

$$\sup_{\theta \in \Theta_0} P(\mathbf{X} \in R_\alpha) = \alpha. \tag{5.1}$$

Throughout this section we assume that the rejection regions are nested in the sense of (2.5) and monotone, that is, for the one sided (2.10)

$$\mathbf{x} \in R_\alpha \text{ and } \mathbf{y} \geq \mathbf{x} \text{ coordinatewise} \Rightarrow \mathbf{y} \in R_\alpha, \tag{5.2}$$

with an analogous requirement for the two sided (2.11). Nested rejection regions are necessary for Theorem 1 to hold, whereas monotonicity is equivalent to (A4). For such tests the acceptance and rejection regions of a fixed α level are connected subsets of $\Re^k$. The points that separate the acceptance from the rejection region are also a connected set. The boundary points can be described by a curve of the form $\{ \mathbf{x} : \psi_\alpha(\mathbf{x}) = 0\}$ where $\psi_\alpha(\mathbf{x})$ depends on α. Note that $\psi_\alpha(\mathbf{x})$ might not have a tractable form. Possibly after reparameterization we can often write,

$$\mathbf{x} \in R_\alpha \Leftrightarrow \psi_\alpha(\mathbf{x}) \geq 0. \tag{5.3}$$

Example 2 Suppose that $X_i \sim N(\theta_i, 1)$ independently and we test $H_0 : \theta_i = 0$, $i = 1, 2, 3$ vs $H_1 : \theta_i \neq 0$ for some i. A test is of the form "Reject if $\mathbf{x} \in R_\alpha = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 - r^2 \geq 0\}$", where r depends on α. The rejection regions are monotone in the two sided sense: if $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{y} = (y_1, y_2, y_3)$, $\mathbf{x} \in R_\alpha$ and $|y_i| \geq |x_i|$, $i = 1, 2, 3$ then $\mathbf{y} \in R_\alpha$. Furthermore they are nested in that a test with smaller α level has a rejection region which is a subset of the rejection region of a test with a larger α level. The boundary points separating the acceptance and rejection regions lie on spheres and have the form $\{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 - r^2 = 0\}$, that is $\psi_\alpha(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 - r^2$.

Rejection regions of most multivariate tests of interest are not as general as (5.3) suggests. Indeed the expressions $\psi_\alpha(\mathbf{x})$ often have the special form

$$\psi_\alpha(\mathbf{x}) = \phi(\mathbf{x}) - c_\alpha \tag{5.4}$$

where $\phi(\mathbf{x})$ is independent of α but c_α is not. If (5.4) holds, we can derive an overall evidential statistic $\phi(\mathbf{x})$ and the testing rules can be written "Reject H_0 if $\phi(\mathbf{x}) \geq c_\alpha$" for some constant c_α, i. e.

$$\mathbf{x} \in R_\alpha \Leftrightarrow \phi(\mathbf{x}) \geq c_\alpha. \tag{5.5}$$

The testing rules, in turn, define a p-value as follows:

$$p(\mathbf{x}) = \inf\{\alpha : \mathbf{x} \in R_\alpha\} = P(\phi(\mathbf{X}) \geq \phi(\mathbf{x})). \tag{5.6}$$

Example 2 (continued) By examining the equation of a sphere we see that $\phi(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2$ is independent of α whereas $c_\alpha = r^2$ depends on α. The Euclidean distance of $\mathbf{x}$ from the origin is a (not calibrated) evidential statistic. The p-value $P(\chi_3^2 \geq x_1^2 + x_2^2 + x_3^2)$, where χ_3^2 is a chi squared random variable with 3 degrees of freedom, is a one-to-one function of $\phi(\mathbf{x})$ and is an evidential statistic in the sense of Section 1.

Remark. If $\alpha_1 \neq \alpha_2$, the curves $\{ \mathbf{x} : \psi_{\alpha_1}(\mathbf{x}) = 0\}$ and $\{ \mathbf{x} : \psi_{\alpha_2}(\mathbf{x}) = 0\}$ might have common points but they do not cross. However if (5.4) holds the curves $\{ \mathbf{x} : \phi(\mathbf{x}) = c_{\alpha_1}\}$ and $\{ \mathbf{x} : \phi(\mathbf{x}) = c_{\alpha_2}\}$ have no common points.

Note that in the one dimensional case, at least when using the Neyman-Pearson approach, rejection regions are nested and monotone and the curve (essentially a point) $\{ \mathbf{x} : \phi(\mathbf{x}) = c_\alpha\}$ that separates acceptance and rejection region defines a function $\phi(\mathbf{x})$ independent of α. Then $\phi(\mathbf{x})$ is unique up to one-to-one transformations and is a one-to-one function of $\mathbf{x}$ (or the p-value). However this is not true in the higher dimensional case. For a given testing rule and data $\mathbf{x}$ there might be more than one curve of the form $\{ \mathbf{x} : \psi_\alpha(\mathbf{x}) = 0 \}$ on which $\mathbf{x}$ lies, hence $P(\phi(\mathbf{X}) \geq \phi(\mathbf{x}))$ is not well defined. The following examples will make clearer the meaning of the function ϕ in a more complicated situation.

Example 3 Suppose that $X_1 \sim f(x_1 \mid \theta_1)$ and we test $H_0 : \theta_1 \leq 0$. If F_0^{-1} is the inverse cumulative distribution function associated with $f(\cdot \mid 0)$, the testing rule "Reject H_0 if $x \in R_\alpha = [\, F_0^{-1}(1-\alpha),\, +\infty\,)$" separates the acceptance and rejection regions by the points $\{ x : x = F_0^{-1}(1-\alpha)\}$ or, equivalently $\{ x : p(x) = \alpha\}$ or any other $\{ x : \phi(x) = c\}$ where ϕ is a one-to-one function of x and c is a one-to-one function of α.

Example 4 Suppose that $X_i \sim f(x_i \mid \theta_i)$ independently and we test $H_0 : \theta_i \leq 0$, $i = 1, 2$.

Consider the testing rule of the form

For $\alpha \in [\, 0, \frac{1}{4}\,]$ reject if $x_1 \geq F_0^{-1}(1-\sqrt{\alpha})$ and $x_2 \geq F_0^{-1}(1-\sqrt{\alpha})$,

For $\alpha \in (\, \frac{1}{4}, \frac{1}{2}\,]$ reject if $x_1 \geq F_0^{-1}(\frac{1}{2})$ and $x_2 \geq F_0^{-1}(1-2\alpha)$,

For $\alpha \in (\, \frac{1}{2}, 1]$ reject if $x_1 \geq F_0^{-1}(1-\alpha)$.

This rule uniquely defines, for any α, the acceptance and rejection regions. However, although for a given α there is a curve $\{ \mathbf{x} : \psi_\alpha(\mathbf{x}) = 0 \}$ separating the regions, the curve cannot be written as $\{ \mathbf{x} : \phi(\mathbf{x}) = c\}$ where ϕ does not depend on α. The point $(x_1, x_2) = (\, F_0^{-1}(\frac{1}{2}), F_0^{-1}(\frac{3}{4})\,)$ lies on more than one curve of the form $\{ \mathbf{x} : \psi_\alpha(\mathbf{x}) = 0 \}$.

Now consider the testing rule

For $\alpha \in [\, 0, 1\,]$ reject if $x_1 \geq F_0^{-1}(1-\sqrt{\alpha})$ and $x_2 \geq F_0^{-1}(1-\sqrt{\alpha})$.

The curve that separates $\Re^2$ into the acceptance and rejection regions is $\{(x_1, x_2) : \min_i x_i = F_0^{-1}(1-\sqrt{\alpha})\}$ or $\{(x_1, x_2) : \min_i p(x_i) = \sqrt{\alpha}\}$. An associated evidential statistic is $\phi(\mathbf{x}) = \min_i p(x_i)$ and the p-value is Tippett' s.

Examining the multivariate p-values derived from rejection regions, one can obtain some interesting results. First we can find the envelopes of the multivariate p-values. That is, for given $\mathbf{x}$, we can find the minimum and maximum p-value that one could report by using various tests, whether they are derived by combining the individual p-values or from some other, distribution specific statistic.

Let $\mathcal{T}$ be the class of all tests that have monotone and nested rejection regions and let $\mathcal{T}_s$ the tests that are, in addition, symmetric in x_i' s. Then, for given $\mathbf{x} \in \Re^k$

$$\inf_{\mathcal{T}} p(\mathbf{x}) = \prod_{i=1}^{k} p(x_i) \tag{5.7}$$

$$\sup_{\mathcal{T}} p(\mathbf{x}) = 1 - \prod_{i=1}^{k} (1 - p(x_i)) \tag{5.8}$$

$$\inf_{\mathcal{T}_s} p(\mathbf{x}) = \min_{1 \le i \le k} p(x_i) \tag{5.9}$$

$$\sup_{\mathcal{T}_s} p(\mathbf{x}) = \max_{1 \le i \le k} p(x_i). \tag{5.10}$$

Verifying (5.7) – (5.10) is a straightforward matter, and follows easily if one draws the pictures associated with the possible shapes of the rejection regions of the tests in $\mathcal{T}$ and $\mathcal{T}_s$. The LHS of (5.7) – (5.10) are the evidential statistics yielding p_F, p_P, p_T and p_M respectively. However, there is no testing rule that achieves the extrema for all $\mathbf{x} \in \Re^k$. If there was, then the statistics themselves would be p-values, hence they would have a uniform distribution, and it is easy to see that it is not the case.

The plethora of tests in the multivariate case, reflecting the wide choice of rejection regions, makes the problem of choosing evidential measures difficult. Even after eliminating the ones that do not obey the axioms of Section 4, there are several plausible functions $\phi(\mathbf{x})$. In the univariate case, appealing to the Neyman-Pearson theory gives reasonable criteria to choose tests and hence associated p-values as good starting points. Other measures of evidence are the Bayesian posterior probabilities of the null hypothesis.

Again, one might ask if decision theoretic tools can help. By evaluating the risks $R_2(\theta, \phi)$ (given by (1.4)) of various $\phi(\mathbf{x})$ as estimators of $I(\theta \in \Theta_0)$, we may hope that we can narrow the choice and arrive at some preferable procedures. Of course, risk considerations cannot be the only criterion. Reasonableness, as formalised in Section 4 is also important. It is well known that, in general, there are procedures that have minimum risk for some parameter values but are otherwise non-optimal. An

attempt to find a single best procedure using only risk considerations is doomed. We may hope, however, that we can find a single procedure, best in terms of risk, in some smaller class of procedures. The following theorem says that this also is a vain hope. The result applies to evidential measures that are p-values derived from the tests in the class $\mathcal{T}$ and shows that for the one-sided hypothesis we cannot find a uniformly best p-value. The theorem is closely related to an analogous result by Birnbaum (1954) for testing. Formally we have:

Theorem 3. For the hypotheses (2.10), let $p(\mathbf{x})$ be any p-value corresponding to a test

$$\mathbf{x} \in R_\alpha \Leftrightarrow \psi_\alpha(\mathbf{x}) \geq 0. \tag{5.11}$$

Then there is a parameter value θ and a p-value $p_\theta(\mathbf{x})$ such that,

$$E_\theta(\, I(\theta \in \Theta_0) - p_\theta(\mathbf{X}))^2 < \; E_\theta(\, I(\theta \in \Theta_0) - p(\mathbf{X}))^2. \tag{5.12}$$

Proof. Observe that $(\, I(\theta \in \Theta_0) - p_\theta(\mathbf{X}))^2$ and $(\, I(\theta \in \Theta_0) - p(\mathbf{X}))^2$ are positive random variables. Using the identity

$$E\; g(X) = \int_0^\infty P(g(X) \geq x)\; dx \tag{5.13}$$

and making the appropriate changes of variables, inequality (5.12) becomes

$$\int_0^1 P_\theta(p_\theta(\mathbf{X}) \leq \alpha)\; (1-\alpha)\; d\alpha < \int_0^1 P_\theta(p(\mathbf{X}) \leq \alpha)\; (1-\alpha)\; d\alpha \tag{5.14}$$

for $\theta \in \Theta_0$, and

$$\int_0^1 P_\theta(p_\theta(\mathbf{X}) \geq \alpha)\; \alpha\; d\alpha < \int_0^1 P_\theta(p(\mathbf{X}) \geq \alpha)\; \alpha\; d\alpha \tag{5.15}$$

for $\theta \notin \Theta_0$. Hence it suffices to find a value of θ and the corresponding $p_\theta(\mathbf{X})$ such that either (5.14) or (5.15) is true. Let $\theta = (\theta_1,\, 0,\, 0, \ldots,\, 0)$ and $p_\theta(\mathbf{X}) = p(X_1)$, the univariate p-value derived from the family of uniformly most powerful unbiased tests for the hypothesis $H_0 : \theta_1 \leq 0$ vs. $H_1 : \theta > 0$. By the definition of most powerful unbiased tests (which have rejection regions of the form $\{\, p(x_1) \leq \alpha\}$), the tests of the form "Reject if $p(\mathbf{x}) \leq \alpha$" must either have a larger rejection probability for a $\theta_1 \leq 0$ or a larger acceptance probability for a $\theta_1 > 0$ (with a strict inequality for some θ_1, unless $p(\mathbf{x}) = p(x_1)$). Hence there exists either a $\theta_1 \leq 0$ such that

$$P_\theta(p(X_1) \leq \alpha) < P_\theta(p(\mathbf{X}) \leq \alpha), \tag{5.16}$$

or a $\theta_1 > 0$ such that

$$P_\theta(p(X_1) \geq \alpha) < P_\theta(p(\mathbf{X}) \geq \alpha). \tag{5.17}$$

Inequalities (5.14) or (5.15) follow immediately from (5.16) or (5.17) respectively. □

In the above theorem the p-value $p(x_1)$ is a rather unreasonable as evidential rule for the hypothesis $H_0 : \theta \leq 0$, but has better risk than other, perhaps more reasonable $\phi(\mathbf{x})$ for some parameter values. A way of eliminating such estimators is to examine the maximum possible risk, that is to apply the minimax criterion. However, in this situation, it turns out that the criterion is not of great help. Similarly to Hwang *et al.* (1992) the unique (under $L_2(\theta, \phi)$) minimax estimator of $I(\theta \in \Theta_0)$ is a rather silly one.

Theorem 4. Suppose the hypothesis are as in (1.1) and Θ_0 and Θ_1 have a common limit point θ_c such that the likelihood $f(\mathbf{x} \mid \theta) = \prod_{i=1}^{k} f(x_i \mid \theta_i)$ is continuous at θ_c. Then the rule $\phi_0(\mathbf{x}) = 1/2$ is unique minimax.

Proof. Consider sequences of points $\theta_{n0} \in \Theta_0$ and $\theta_{n1} \in \Theta_1$ such that $\lim_{n\to\infty} \theta_{nj} = \theta_c$, $j = 0, 1$ and a sequence of priors putting a mass $1/2$ to each of the points θ_{n0} and θ_{n1}. It is straightforward to see that the corresponding sequence of Bayes rules is

$$\phi_n(\mathbf{x}) = \frac{1}{1 + \dfrac{f(\mathbf{x} \mid \theta_{n0})}{f(\mathbf{x} \mid \theta_{n1})}} \tag{5.18}$$

and by the continuity of the likelihood $\lim_{n\to\infty} \phi_n(\mathbf{x}) = 1/2$. By the Dominated Convergence Theorem the limiting Bayes risk is equal to $1/4$, hence by applying Theorem 18, p. 350 of Berger (1985), we conclude that $\phi_0(\mathbf{x})$ is minimax. Uniqueness follows from the convexity of the loss $L_2(\theta, \phi)$. □

Another criterion which has proved useful in eliminating sub-optimal procedures is admissibility. Though there are admissible rules that are certainly unreasonable, the use of inadmissible rules is rather undesirable since they can be dominated by others for all parameter values. In the next section a criterion for the admissibility of evidential rules is provided.

6. Characterizing Admissible Rules

We turn our discussion to the admissibility of evidential procedures with respect to the squared error loss $L_2(\theta, \phi)$ of (1.3). The general theorem is, that under some regularity conditions that we will make precise, Bayes rules are admissible procedures. Earlier work (Hwang *et al.* 1992) showed that the generalised Bayes rules form a complete class, but it was assumed that $\Theta \subset \Re$ and the family of distributions was a canonical exponential family on $\Re$. In this section we extend this result to multidimensional problems. Although we will only apply the complete class results to the combining of experiments problem, the results can be applied in other settings.

Note that if admissibility is the only criterion, this complete class result almost rules out the consideration of p-values based on the evidential statistics of Table 1 because they are not Bayes rules. This is somewhat contrary to the one dimensional situation (Hwang *et al.* 1992), where p-values are generalised Bayes rules in the one-sided problem, hence admissible. In higher dimensions, for the problem of one-sided testing, there are statistics based on p-values which are generalised Bayes, such as the product of p-values. It turns out that this is an admissible procedure. Based on admissibility, rather than using the Fisher combination as the basis for constructing a p-value, one could use the product itself as a measure of the evidence that has been accumulated against the null hypothesis. The problem is, as we have already seen, that the values of the product are in general too small to be intuitively acceptable since the prior probability of the null is small.

We shall consider the testing problem for $\Theta \subset \Re^k$. We assume that the model under consideration is a canonical exponential family, with density (with respect to a measure ν) given by

$$f(\mathbf{x} \mid \theta) = \exp\{\theta \cdot \mathbf{x} - \psi(\theta)\} \qquad \theta \in N_\nu \tag{6.1}$$

where

$$N_\nu = \{\theta: \int \exp\{\theta \cdot \mathbf{x}\}\nu(d\mathbf{x}) < \infty\} \tag{6.2}$$

$$\psi(\theta) = \log \lambda_\nu(\theta) \tag{6.3}$$

with

$$\lambda_\nu(\theta) = \int \exp(\theta \cdot \mathbf{x})\, \nu(d\mathbf{x})\,. \tag{6.4}$$

Note that $\lambda_\nu(\theta)$ is the Laplace transform of the dominating measure ν, with θ as the parameter of the transform. One may also define $\lambda_\nu(\theta)$ as a Laplace transform, however now, integrating with respect to $\nu(d\theta)$. The recognition that $\lambda_\nu(\cdot)$ is a Laplace transform simplifies the proof of the construction of the complete class theorem. Using this approach, one may apply a variety of results on the continuity of the Laplace transform. For more details on these continuity results see Brown (1976, Chapter 2). The results will be applied in the special case of the problem at hand. The main ideas of the proof go back to the pioneering work of Farrell (1968).

It is shown below that the rules in the complete class are essentially generalized Bayes rules, after allowance for truncation. We define $\mathcal{C}$ to be a *truncation set* for a function $\phi(\mathbf{x})$ if $\phi(\mathbf{x}) = 0$ for $\mathbf{x} \notin \mathcal{C}$. For more on truncation see Stein (1956) and Farrell (1968). In the following we consider the equalities to be equalities almost everywhere (with respect to the appropriate measure).

Theorem 5. Let ϕ be an admissible estimator of $I(\theta \in \Theta_0)$ under the loss function $L_2(\theta, \phi)$ and $\mathcal{C}$ be a truncation set for ϕ such that for all $\mathbf{x} \in \mathcal{C}$, $0 < \phi(\mathbf{x}) < 1$. Then there exist σ–finite measures π_0 on Θ_0 and π_1 on Θ_1 such that

$$\int e^{\mathbf{x} \cdot \theta - \psi(\theta)} [\pi_0(d\theta) + \pi_1(d\theta)] = 1 \qquad \forall\, \mathbf{x} \in \mathcal{C} \tag{6.5}$$

and ϕ is finite and given by

$$\phi(\mathbf{x}) = \frac{\int_{\Theta_0} f(\mathbf{x} \mid \theta)\, \pi_0(d\theta)}{\int_{\Theta_0} f(\mathbf{x} \mid \theta)\, \pi_0(d\theta) + \int_{\Theta_1} f(\mathbf{x} \mid \theta)\, \pi_1(d\theta)}. \tag{6.6}$$

Proof: Suppose ϕ is an admissible rule. From Brown (1986, Theorems 4.A.7 and 4.A.12) there exists a sequence of finite priors G_n concentrated on finite subsets such that the Bayes rule $\phi^{G_n}(\mathbf{x})$ converges to $\phi(\mathbf{x})$ in the weak* topology. The special case of $\phi \equiv 0$ is obvious. Assume $E_\theta \phi(\mathbf{X}) > 0$. By the dominated convergence theorem it follows that $\phi^{G_n}(\mathbf{x}) > 0$ with positive measure for n sufficiently large. Define

$$H_{in}(d\theta) = \frac{e^{-\psi(\theta)}\, G_n(d\theta)}{\int_{\Theta_0} e^{-\psi(\theta)}\, G_n(d\theta)}, \quad \theta \in \Theta_i, \quad i = 0, 1. \tag{6.7}$$

Let $H_n = H_{0n} + H_{1n}$. Note that H_n is a finite measure. By Lemma 7.17 of Brown (1986) there exists a subsequence $H_{n'}$, limiting finite measure H and a closed convex set $\mathcal{C}$ on $\bar{\Theta}$, such that for $i = 0, 1$, as $n' \to \infty$

$$\lambda_{H_{in'}}(\mathbf{x}) \to \lambda_{H_i}(\mathbf{x}) \qquad \mathbf{x} \in \mathcal{C}^0 \tag{6.8}$$

$$\lambda_{H_{in'}}(\mathbf{x}) \to \infty \qquad \mathbf{x} \notin \mathcal{C}. \tag{6.9}$$

Note $H(\Theta) = H_0(\Theta_0)$ (since $H_{0n}(\Theta_0) = 1$). Now define

$$\phi_n(\mathbf{x}) = \frac{\lambda_{H_{0n}}(\mathbf{x})}{\lambda_{H_{0n}}(\mathbf{x}) + \lambda_{H_{1n}}(\mathbf{x})}. \tag{6.10}$$

It follows by construction that $\phi_n \to \phi$ for all $\mathbf{x} \in \mathcal{C}$.

The first part of the theorem follows from the fact that Bayes estimates are admissible and the fact that for L_2 loss, if a rule ϕ' is as good as ϕ and ϕ has a truncation set $\mathcal{C}$, then $\phi'(\mathbf{x}) = \phi(\mathbf{x})$ for $\mathbf{x} \notin \mathcal{C}$. □

Remark 1. The results hold if ν is a counting measure.

Remark 2. The truncation set may be identified with the convex acceptance region, as studied by Birnbaum (1955) and Stein (1956), in the case where Θ_0 is a bounded set in $\Re^k$. It may be shown that any nonrandomised test with convex acceptance region is admissible. In fact, when $\Theta_0 = \{\theta_0\}$ is simple, such tests form a minimal complete class. These facts link the optimality results of testing (i.e. L_1-loss) to those in evidence assessment (i.e. L_2-loss). That is, the truncation sets in the complete class theorem above are to be identified with the convex acceptance regions in simple hypothesis testing.

Now consider the problem of testing k separate independent hypothesis, that is, testing (2.9) where $\Theta_0 = \bigcap_i \Theta_{0i}$. It is important to note that the null hypothesis states that all $\theta_i \in \Theta_{0i}$ simultaneously. Therefore for each event $\theta_i \in \Theta_{0i}$ one may define a parameter $I(\theta_i \in \Theta_{0i})$ which is to be estimated. Hence the parameter of interest for H_{0i} $(i = 1, \ldots, k)$ is equal to $\prod_{i=1}^{k} I(\theta_i \in \Theta_{0i})$. Following the development above and in Hwang *et al.* (1992) it follows that the loss function will be

$$L_2(\theta, \phi) = \left(\prod_{i=1}^{k} I(\theta_i \in \Theta_{0i}) - \phi(\mathbf{x}) \right)^2 . \tag{6.11}$$

Under the conditions of the complete class theorem above it follows that the admissible estimators are the generalized Bayes rules. Hence, if the prior for $\theta = (\theta_1, \ldots, \theta_k)$ is a product of the priors for $\pi(\theta_i)$ $(i = 1, \ldots, k)$, it follows that the product of the individual generalized Bayes rule for H_{0i} form the complete class for the multiple hypothesis H_0.

Of course admissibility of an evidential rule implies that there is no other rule that dominates it for all parameter values, hence it rules out any complete risk ordering of admissible estimators. It is, however, necessary to examine numerically the risks of various rules as functions of the parameters in order to gain some feeling as to whether some estimators are more satisfactory than others. Figure 1 shows some simulation results for the normal case and the one sided null hypothesis. The plot shows the risks of the Fisher and Tippett p-values, as well as the Fisher statistic, (i. e. the product of the individual p-values) and the impartial Bayes rule $\phi_L(\mathbf{x})$. It can be seen that the risks of Tippett's and Fisher p-values are virtually identical. The Fisher statistic and the Bayes rule have relatively small risk in the alternative, but they do not perform satisfactorily in the null. Furthermore, the differences in the risks in the null and in the alternative is large. This can be explained by the fact that they are both small, hence they perform well when they estimate the indicator function when it equals zero but badly when the indicator function equals one.

7. Discussion

Evidence in the multivariate case seems to be considerably harder to quantify than in the univariate case. In some sense there are too many criteria and too many procedures to choose from. We would like to use statistics that both conform to our intuition and perform satisfactorily from a decision theoretical point of view. The axioms formulated here give a quantitative description of our intuition whereas minimaxity and admissibility results can provide evaluation tools. However, there does not seem to be a clear connection between the two approaches. Rules that perform well with respect to the evidential axioms may not necessarily do so under risk considerations.

It seems that the omnibus p-values derived from the various combination rules of the individual p-values are not posterior rules under any prior. Tippett's and maximum p-values are not smooth enough (they are not differentiable as functions of the data) to be Bayes or generalised Bayes rules, so

they are ruled out immediately. Other rules such as normal, sum or Fisher also seem unlikely to be generalised Bayes. This should not come as a surprise, since such combinations were constructed in rather ad hoc, though heuristically intuitive ways. However the heuristic way of constructing them implies, as a result, that they tend to follow the evidence desiderata.

On the other hand, Bayes rules that are optimal using decision theoretical criteria need not be evidential, at least in the sense that was described in the paper. Statistics such as the product of p-values are too small to be satisfactory whereas attempts to construct "impartial" Bayes rules make some difficulties apparent. The dimension of the problem affects the posterior probabilities in a crucial way. In order to obtain posterior probabilities that do not bring evidence against the null hypothesis too often, we must put a large prior mass on Θ_0, so in higher dimensions there is a built in bias against the null.

Each particular problem will have considerations that are important. If a well calibrated, intuitive evidential statistic is desired, then a p-value such as Fisher or Tippett should be used. They both seem to behave reasonably for different parameter values, both in the null and the alternative. This is perhaps closely connected with the convexity of the regions of the tests from which they are derived. Under some regularity conditions, tests with convex acceptance regions form complete class with respect to the traditional decision theoretical loss for tests (i. e. the absolute error loss $L_1(\theta, \phi)$), and it seems that this optimality carries through to p-values as evidential rules. Other p-values do not to satisfy all axioms, so it should not come as a surprise if, at least for some data values, they give evidence contrary to our intuition.

If we are not concerned with evidence as formalised above, but with the estimation of accuracy in testing, such as expressed by the indicator function, then generalised Bayes rules are clear winners. The impartial rule which gives equal mass to the null and the alternative satisfies the admissibility criterion, as well as our idea of impartiality. Of course, if prior information is available and we want to use it, then there is no reason to be impartial and proper Bayes posterior probabilities can be reported as evidential measures.

It seems, however, that calibration of evidence is more important than optimality using decision theoretical criteria, which necessitates the existence of some axioms that evidential rules should satisfy. Minimaxity proves to be useless, whereas admissibility does not guarantee that only reasonable estimators are optimal in that criterion. Furthermore the fact that a rule is inadmissible does not imply that an estimator that has smaller risk is easy to find or to use. Numerical evidence shows that the risk of the admissible rules can be quite high compared with the risk of the reasonable evidential rules such as the Fisher and Tippett p-values. The simulation results also suggest that these p-values have risk close to the minimax risk. The relatively good decision theoretical performance coupled with their intuitively appealing behaviour justifies their use as rules of evidence.

Appendix

Lemma A.1 Let $X \sim f(x \mid \theta)$, and suppose we test $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ where the support of $f(x \mid \theta)$ does not depend on θ. Let $\pi(\theta)$ be a prior supported on $\Theta_0 \cup \Theta_1$ for which $\pi(\Theta_0) > 0$ and $\pi(\Theta_1) > 0$, and which results in a marginal distribution $m_\pi(x) < \infty$ for every x.

i) If there is a value a for which

$$\lim_{x \to a} \frac{f(x \mid \theta_0)}{f(x \mid \theta_2)} = 0 \quad \text{and} \quad \lim_{x \to a} \frac{f(x \mid \theta_1)}{f(x \mid \theta_2)} \neq 0 \tag{A.1}$$

monotonically for all $\theta_0 \in \Theta_0$, all $\theta_1 \in \Theta_{\pi_1} \subseteq \Theta_1$ where $\pi(\Theta_{\pi_1}) > 0$ and some θ_2, then

$$\lim_{x \to a} P(\theta \in \Theta_0 \mid x) = 0. \tag{A.2}$$

ii) If there is a value b for which

$$\lim_{x \to b} \frac{f(x \mid \theta_1)}{f(x \mid \theta_2)} = 0 \quad \text{and} \quad \lim_{x \to b} \frac{f(x \mid \theta_0)}{f(x \mid \theta_2)} \neq 0 \tag{A.3}$$

monotonically for all $\theta_1 \in \Theta_1$, all $\theta_0 \in \Theta_{\pi_0} \subseteq \Theta_0$ where $\pi(\Theta_{\pi_1}) > 0$ and some θ_2, then

$$\lim_{x \to b} P(\theta \in \Theta_0 \mid x) = 1. \tag{A.4}$$

Proof: The proof follows quickly from Lebesgue's monotone convergence theorem by writing

$$\frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)} = \frac{\int_{\Theta_0} \frac{f(x \mid \theta)}{f(x \mid \theta_2)} \pi(\theta)\, d\theta}{\int_{\Theta_1} \frac{f(x \mid \theta)}{f(x \mid \theta_2)} \pi(\theta)\, d\theta}. \tag{A.5}$$

Taking limits inside the integral yields the desired result. □

REFERENCES

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis,* Second Edition. New York: Springer-Verlag.

Berger, J. O. and Delampady, M. (1987). Testing precise hypothesis (with discussion). *Statist. Sci.* **2**, 317-352.

Birnbaum, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49**, 559-574.

Birnbaum, A. (1955). Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests. *Ann. Math. Statist.* **26**, 21-36.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* **57**, 269-306.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families.* IMS Monograph Series, Institute of Mathematical Statistics, Hayward, CA.

Casella, G. and Berger, R. L. (1987). Reconciling evidence in the one-sided testing problem (with discussion). *J. Amer. Statist. Assoc.* **82**, 106-111.

Farrell, R. H. (1968). Towards a theory of generalized Bayes tests. *Ann. Math. Statist.* **39**, 1-28.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*, Fourth Edition. Edinburgh: Oliver and Boyd.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* New York: McGraw Hill.

Hedges, L. V. and Olkin, L. (1985). *Statistical Methods for Meta-Analysis.* San Diego: Academic Press.

Hwang, J. T., Casella, G., Robert, C., Wells, M. T., and Farrell, R. H. (1992). Estimation of accuracy in testing. *Ann. Statist.* **20**, 490-509.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789-827.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, Second Edition. New York: John Wiley.

Marden, J. I. (1991). Sensitive and sturdy p-values. *Ann. Statist.* **19**, 918-934.

Moreno B., E. and Cano S., J. A. (1989) Testing a point null hypothesis: Asymptotic robust Bayesian analysis with respect to the priors given on a subsigma field. *Int. Statist. Rev.* **57**, 221-232.

Robinson, G. K. (1979). Conditional properties of statistical procedures. *Ann. Statist.* **7** 742-755.

Royall, R. (1986). The effect of sample size on the meaning of significance tests. *Amer. Statist.* **40**, 313-315.

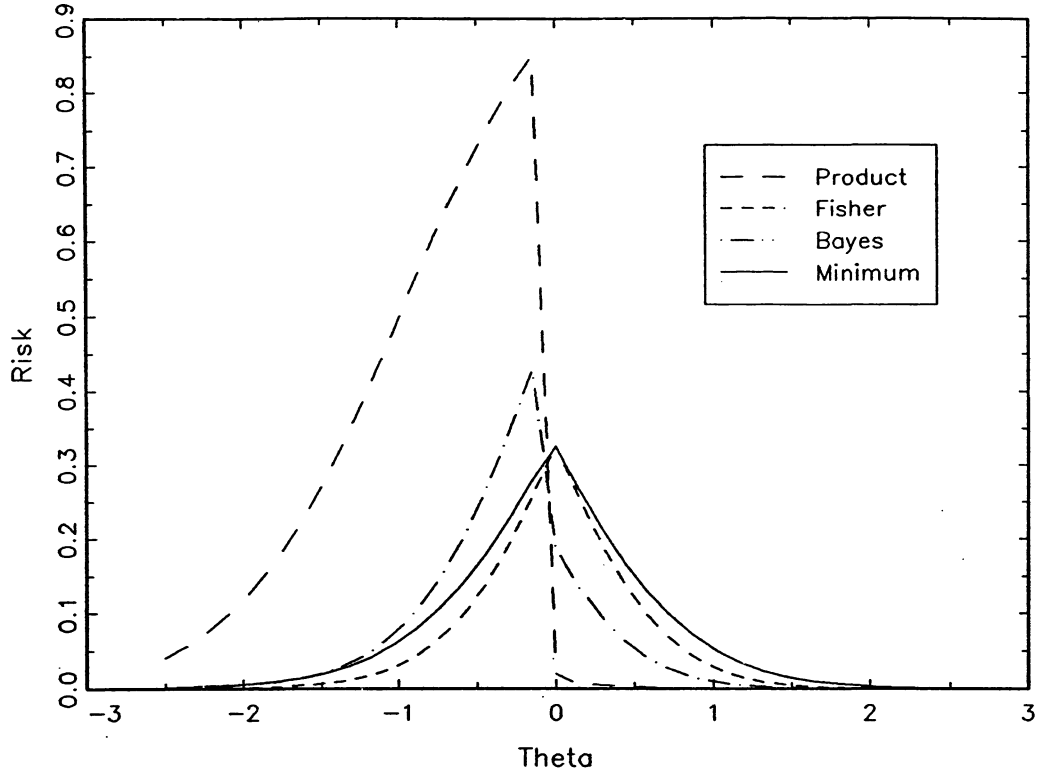
Tippett, L. H. C. (1931). *The Methods of Statistics*, First Edition. London: Williams and Norgate.

Stein, C. (1956). The admissibility of Hotelling's T^2 test. *Ann. Math. Statist.* **27**, 616-623.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin* **48**, 156-157.

Figure 1: Simulated risks of the Fisher statistic (Product), the Fisher p-value (Fisher), the impartial Bayes rule $\phi_L(\mathbf{x})$ (Bayes) and Tippett p-value (Minimum) for a normal distribution and one-sided hypothesis $H_0 : \theta_i \leq 0$, i = 1, 2, 3, 4. The true parameter values have the form (θ, θ, θ, θ) for (a) and (0, 0, θ, θ) for (b). The x-axes of the plots represent the values of θ.

(a)



(b)